Bacterial whole-genome sequences: minimal information and strain availability

The number of prokaryotic genome sequences that has been determined has increased steadily over the past five years, and genomes are now published almost on a weekly basis (Ussery, 2004). In November 2001, Ward and colleagues wrote a letter to *Nature* (Ward et al., 2001) in which they pointed out that many prokaryotic strains used for genome-sequencing projects are poorly documented and not publicly available. They proposed three standards that should be adopted by the entire community: (i) sequenced strains should be deposited in at least two major public biological resource centres (BRCs), (ii) lists and databases should include the name of the sequenced strain, its origin and the associated culture collection accession numbers, and (iii) the type strain of a species should be sequenced unless other factors make this inappropriate.
We assessed to what extent these recommendations have been followed for bacterial genome sequences deposited in the GenBank database (both the 'complete' and the 'in progress' databases found at http://www.ncbi.nlm.nih.gov/ genomes/MICROBES/) and unfinished and/or unpublished bacterial genome sequences available through the websites of The Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/Microbes), The Institute for Genomic Research (TIGR) (http://www.tigr.org/tdb/mdb/mdbinprogress.html), the DOE Joint Genome Institute (JGI) (http://www.jgi.doe.gov/JGI_microbial.html/index.html) and Genoscope (http://www.genoscope.cns.fr/externe/English/Projets/). We did not include projects concerning uncultured environmental or symbiotic strains.

A total of 323 bacterial genome sequences were retrieved from the above-mentioned databases. In 31 of these projects, there was not a single indication of the identity of the strain being sequenced (i.e. no strain number was provided; e.g. *Actinobacillus pleuropneumoniae*, GenBank accession nos NC_004130 and NC_004427). In addition, ten species names that are used have not been validly described (e.g. *Bacteriovorax marinus* and *Dechloromonas aromatica*), although some of them have been used extensively in the literature (e.g. *Haemophilus somnus*). It should be noted that all sequences for which there is no clear strain designation are at present unpublished, while this is also true for most (8 out of 10) of the sequences associated with invalid species names.

The bacterial isolates for which the complete genome sequence was determined belong to 235 different species. Sixty-five sequenced strains (20·1%) are type strains. Thirty-two species are represented by more than one sequenced strain; for only seven of these (21·9%) was the type strain among the isolates being sequenced. However, as the underlying reason for the initiation of many genome-sequencing projects may be the presence of special properties in the strain proposed for sequencing (e.g. highly virulent isolates, isolates capable of degrading various compounds, etc.), this should not come as a surprise.

Approximately half of the strains (163, 50·5%) have been deposited in at least one major public BRC, indicating that 160 bacterial strains (49·5%) are currently not publicly available. Ninety-seven bacterial strains (30·0%) have been deposited in at least two major public BRCs. Of the 160 strains not deposited in any major public BRC, the genome sequence has not been published yet for 64 (40·0%). Surprisingly, for only 50 of all 160 sequenced strains (31·3%) deposited in at least one BRC is the culture collection number used as primary designation.

The data presented here demonstrate that the standards proposed by Ward et al. (2001) have been largely ignored. However, considering the amount of money and effort spent in determining and annotating the genome sequence, and in post-genomic studies, the scientific community should continue to demand deposition of sequenced strains in internationally recognized BRCs.

Tom Coenye and Peter Vandamme

Laboratorium voor Microbiologie, Universiteit Gent, B-9000 Gent, Belgium

Correspondence: Tom Coenye (Tom.Coenye@UGent.be)


DOI 10.1099/mic.0.27307-0